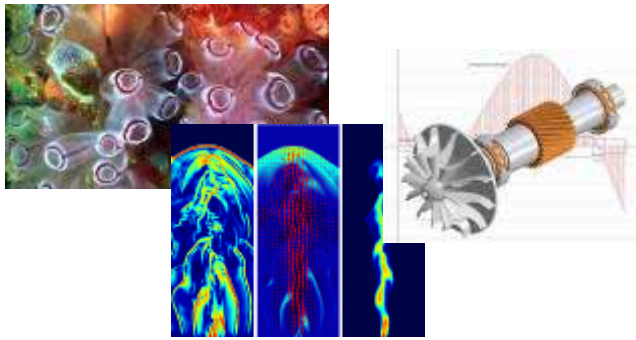
EGEE-II
Bioinformatics Activity

Dr Christophe Blanchet
EGEE Bioinformatics Activity Leader
CNRS IBCP, Lyon, France
Christophe.Blanchet@ibcp.fr

The EU EGEE grid project

- **EGEE-I**
  - 1 April 2004 – 31 March 2006
  - 71 partners in 27 countries, federated in regional Grids
- **EGEE-II**
  - 1 April 2006 – 31 March 2008
  - 91 partners in 32 countries
  - 13 Federations
- **Objectives**
  - Large-scale, production-quality infrastructure for e-Science
  - Improving and maintaining "gLite" Grid middleware
  - Attracting new resources and users from industry as well as science

Size of the infrastructure (Sept. 2006 @ EGEE06):
192 sites in 40 countries
~25 000 CPU
~ 5 PB disk, + tape MSS



30+ Grid Projects @ EGEE'06

Enabling Grids for E-sciencE

Businesses @ EGEE06

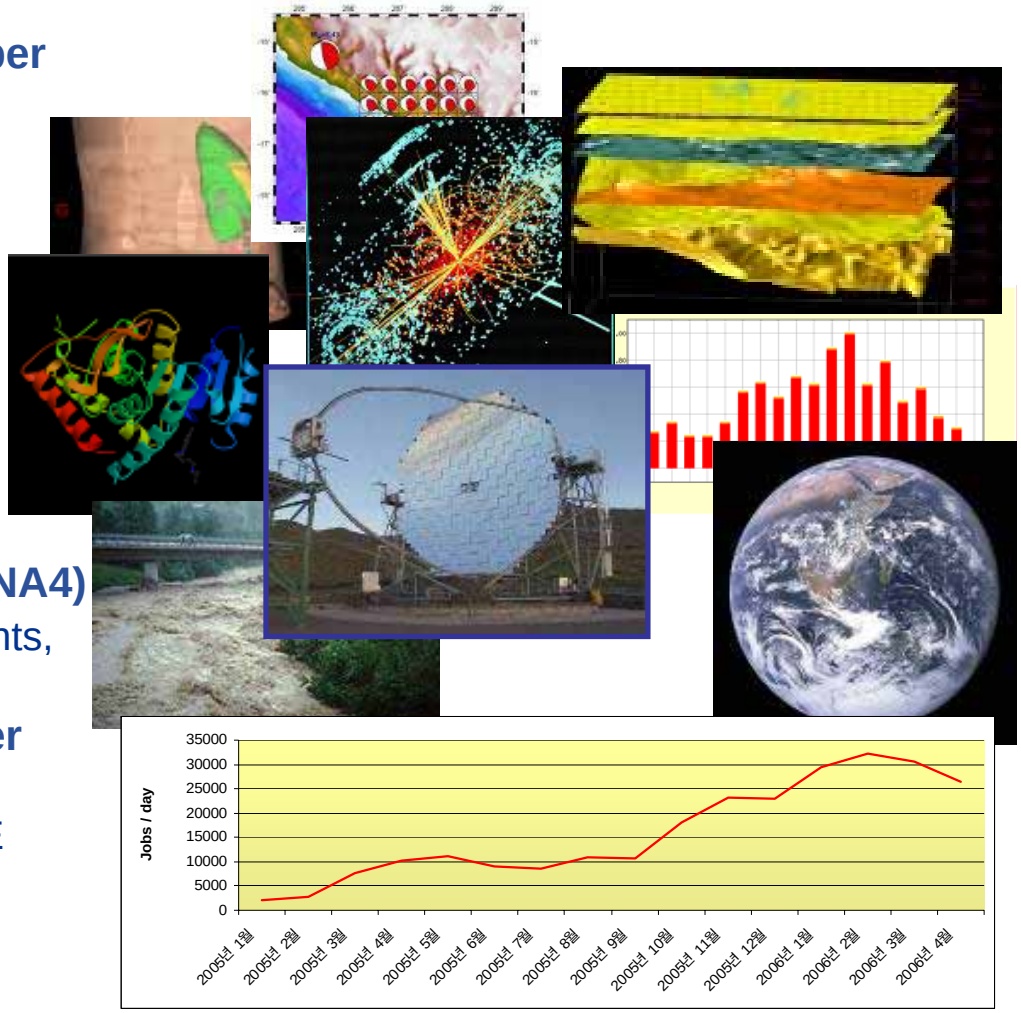
Atos Origin
CS
DATAMAT A Finmeccanica Company
GlobalVillage
Media Sponsor GRIDtoday
gridwise tech
hp invent
intel Leap ahead
intellect
kite knowledge · innovation · technology · enterprise

Capitalising on e-Science to make e-Business

maat
metaware
GRUPPO MPS CONSORZIO OPERATIVO GRUPPO MPS
NICE Networking Engineering Communication nice-italy.com
ontoprise know how to use Know-how
ORACLE
Platform
PRICEWATERHOUSECOOPERS
Qi3
SIEMENS
SOURCING & OUTSOURCING SERVICES LTD
Sun microsystems
TOTAL
T-Systems

Applications

- **Many applications from a growing number of domains**
  - Astrophysics
  - Computational Chemistry
  - Earth Sciences
  - Financial Simulation
  - Fusion
  - Geophysics
  - High Energy Physics
  - Life Sciences
  - Multimedia
  - Material Sciences
  - ...
- **Application Identification and Support (NA4)**
  - 25 countries, 40 partners, 280+ participants, 1000s of users
- **Support the large and diverse EGEE user community:**
  - **Promote dialog**: Users' Forums & EGEE Conferences
  - **Technical Aid**: Porting code, procedural issues
  - **Liaison**: Software and operational requirements



"Biomedical" Virtual Organization

- **Biomed VO Management**
  - Leader: V. Breton
  - Deputies: C. Blanchet & J. Montagnat
  - ~80 participants
- **Three active subgroups**
  - Bioinformatics (C. Blanchet)
    - Details in next slides
  - Drug discovery (V.Breton)
    - Successful runs for malaria and avian flu virus.
    - Similar work to be done for neglected diseases in EGEE-II.
    - WISDOM: 1 October - 1 December, 500 CPU-years, 5 TB, Discussions underway for finalizing docking targets
  - Medical imaging (J. Montagnat)
    - Kickoff meeting on July 12 in Sophia Antipolis
    - Three application services offered from partners (MDM, Moteur, P-grade)
    - 6 applications from EGEE, 5 new in EGEE-II
- **Infrastructure (Dec 2006)**
  - Computing: 113 CEs, ~15,000 CPUs
  - Storage: 107 SEs, ~3,5 TB
  - ~1000 jobs/day

Bioinformatics Activity

- **The bioinformatics sector targets gene and protein analysis, for example genomics, proteomics and phylogeny; but also system biology, genetic linkage, genetic demographic model, …**

- **Integrate biological data and tools**
  - Select relevant applications with real grid add-value
  - Define and prioritize their requirements
  - Give feedback about satisfaction with middleware components
- **Promote Dialog**
  - Internal Bioinformatics meetings
  - Participation to other EGEE activity meetings
  - Scientific dissemination in national and international conferences
  - Collaboration with related projects:
    - EU-EMBRACE, EU-EELA, EU-BIOINFOGRID, EU-ETICS, SwissBioGrid.
    - Joint meetings with related projects
- **Grid expertise**
  - Consulting about EGEE grid platform
  - Help on porting bioinformatics applications
  - Train users and developers
  - Support, helpdesk
- **Deploy applications on the production platform**
  - 10 applications
  - 4 applications from EGEE, 6 new ones in EGEE-II
  - Training, collaboration with Regional Operation Center,
  - Add new resources: hardware and human
  - Give feed-back of services use.

NoE EMBRACE (EU FP6)

- **« A European Model for Bioinformatics Research and Community Education »**
  - simplify and standardize the way in which biological information is served to the researchers who use it.
  - Integrating biological data and bioinformatics tools in grid
- **Network of Excellence (2005-2010)**
  - From Feb 1st, 2005
  - partners: EBI (PI), EMBL, SIB, CNRS, MPI_MG, INRA, ITB CNR, CNB, ...
- **Funded by the European Union (EU-FP6, LHSG-CT-2004-512092)**
  - EMBRACE uses a test problem driven development method. The services will be developed through a set of test problems, which will use tasks from real biological research, designed to stretch the system in critical ways

Bioinformatics Applications

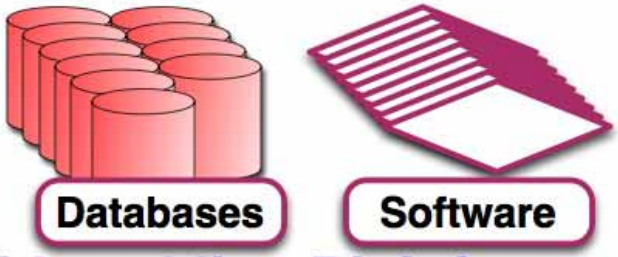
GPS@	CNRS IBCP	Christophe Blanchet Christophe.Blanchet@ibcp.fr	Prototype	http://gpsa-pbil.ibcp.fr/
SPLATCHE	Univ. Bern	Nicolas Ray nicolas.ray@zoo.unibe.ch	Production	**http://cmpg.unibe.ch/software/splatche/**
Large-scale Pathway Analysis	MPI-MG	Ralf Herwig herwig@molgen.mpg.de	Porting	
bioDCV	INFN,ICTP	Cesare Furlanello furlan@itc.it	Production	**http://biodcv.itc.it/**
Phylojava	CNRS	Manolo Gouy mgouy@biomserv.univ-lyon1.fr	Porting	**http://pbil.univ-lyon1.fr/software/phylojava/phylojava.html**
BiG	UPV	Ignacio Blanquer iblanque@dsic.upv.es	Porting	
Superlink-online	TAU	Mark Silberstein marks@techunix.technion.ac.il	Feasibility	**http://bioinfo.cs.technion.ac.il/superlink-online/**
3DEM	CNB/CSIC	Jose-Maria Carazo carazo@cnb.uam.es	Porting	**http://3dem.ucsd.edu/**
CAST	UCY	George Tsouloupas georget@ucy.ac.cy	Feasibility	
swissPIT	SIB/CSCS	Patricia Hernandez Patricia.Hernandez@isb-sib.ch	Feasibility	

GPSA: Bioinformatics Grid Portal

- **Scientific objectives**
  - –Molecular Bioinformatics: protein sequence analysis
  - –Analyse data from high-throughput Biology: complete genome projects, EST, complete proteomes, structural biology, ….
  - –Integration of biological data and tools
- **Method**
  - –Provide Biologists with an usual Web interface: NPS@
    - NPS@ Web portal online since 1998
    - 46 tools & 12 updated databases
    - + 9,000,000 jobs & 5,000 jobs/day
  - –Ease the access to updated databases and algorithms.
  - –Protein databases are stored on grid storage as flat files.
  - –Legacy bioinformatics applications
    - Wrapping usual binary in grid environment
    - transparent remote access with local filesystem
  - –Display results in graphical Web interface.
- **Status:** Prototype

GPSA Results

- **Grid-enabled bioinformatics resources**
  - –9 algorithms
  - –3 protein databases
- **Bioinformatics** descriptors
  - –XML framework, WSRF
- **Encryption system:** EncFile
  - –Security: AES, Key sharing, M-of-N
- **Transparent access to grid files:** Perroquet
- **http://**gpsa-pbil.ibcp.fr

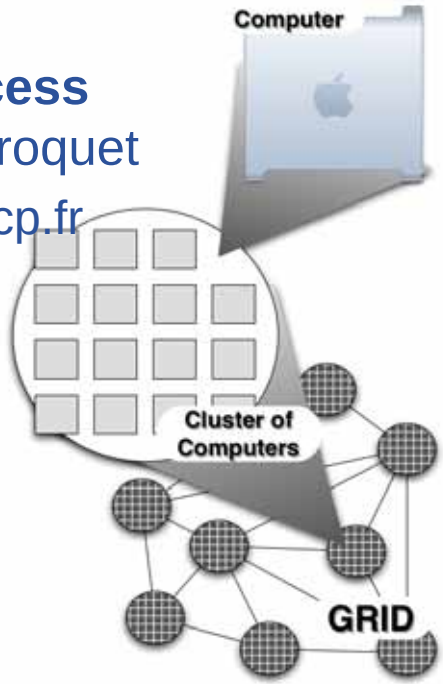


Grid-enabling Bioinformatics

Legacy Bioinformatics Applications :
* Wrapping tools with XML descriptors
* BLAST, SSEARCH, FASTA, ClustalW, MultAlin, PattInProt, ...

Distributed databases:
* **Encrypting Data** with EncFile system
* **Swiss-Prot:**
lfn:/grid/biomed/db/swissprot/last/sprot.fas
Symlink to *(...)/swissprot/50/4/sprot.fas*
* **TrEMBL:**
lfn:/grid/biomed/db/trembl/33/4/sptr.fas

Details : http://gbio.ibcp.fr



SPLATCHE (1)

SPatiaL And Temporal Coalescences in Heterogeneous Environment

http://cmpg.unibe.ch/software/splatche

- **Scientific objectives**

  Study **human evolutionary genetics** and answer questions such as the **geographic origin** of modern human populations, the **genetic signature** of expanding populations, the **genetic contacts** between modern humans and Neanderthals, and the expected null distributions of genetic statistics applied on **genome-wide data sets**.

- **Method**

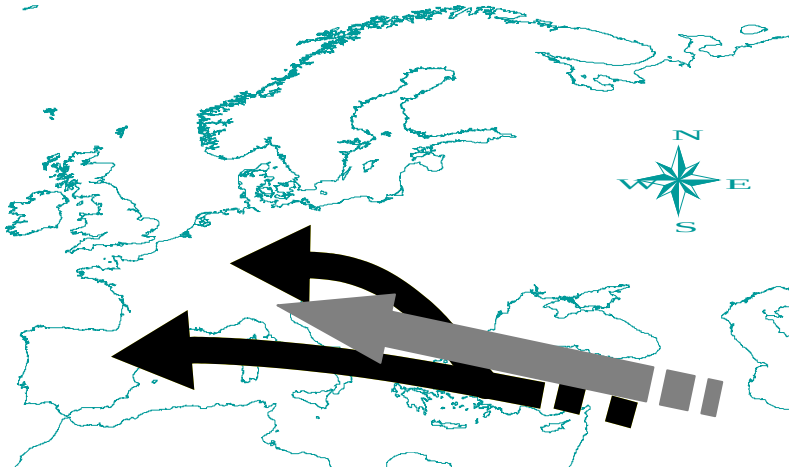
  Simulate the past demography (growth and migrations) of human populations into a geographically realistic landscape, by taking into account the spatial and temporal heterogeneity of the environment.

  Generate the molecular diversity of several samples of genes drawn at any location of the current human's range, and compare it to the observed contemporary molecular diversity.

  SPLATCHE uses a region sampling Bayesian framework that requires10^5 independent demographic and genetic simulations.

SPLATCHE (2)

- **Comparison of 4 different demographic models of human evolution, using a new set of nuclear markers**
- **Results**
  - 40-min jobs are a good compromise between # of CPUs and # of jobs
  - 2 mio simulations 4'000 jobs
  - About 80 CPU-days per try
  - 2 tries had 0% job failure
  - 2 other tries had about 2-3% job failures



Expansion of first farmers 8,000-3,000 years ago

Expansion of early modern Europeans 45,000-30,000 years ago

Simulation replicates on the grid

In black: relative proportion of genes brought by first farmers in various European regions (maximum=32%)

Probability distribution of demographic / genetic parameter of interest (admixture, migration, growth rates, etc…)

Large-Scale Pathways Analysis (1)

QuickTime?et un
d?ompresseur TIFF (LZW)
sont requis pour visionner cette image.

Given the metabolic network of an organism, the application will screen high-throughput data derived from Protein-Protein Interaction and DNA array experiments in several conditions (for example a disease and a control condition) and identify important nodes in the network that show significant concentration changes:

- Pure Metabolic Model: e.g. from KEGG (>1400 reactions)
- Metabolic & Signal-Transduction pathways: e.g. from Reactome (>1500)

Large-Scale Pathways Analysis (2)

Applications

Cancer application

Melanoma celllines:
[**primary tumor** vs. **metastases** (tumor progression) vs. **control**]

Homepage of ESBIC-D (EU project):
http://pybios.molgen.mpg.de/ESBIC-D

Type-2 diabetes

Mouse model (NZO):
[**standard diet** vs. **high fat diet**]

Nutrigenomik (BioProfile – BMBF – Germany):
http://www.molgen.mpg.de/~lh_bioinf/projects/Nutrigenomik/

BioDCV

Application for analysis of microarray and proteomics data with Support Vector Machine (SVM) classifiers; with IFOM-FIRC -- BICG AIRC project

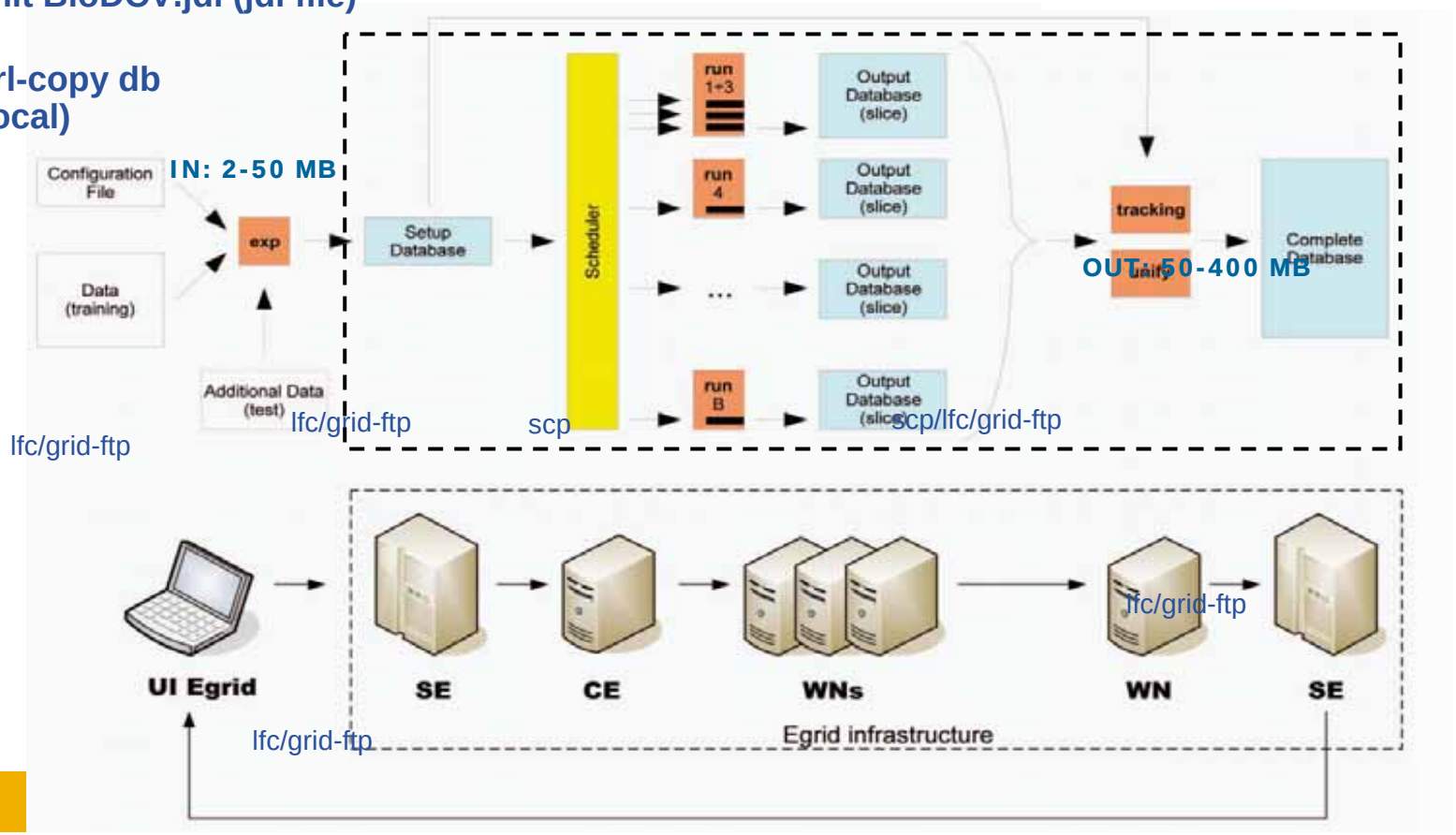
Standard LCG user interface commands are used to transfer

a. *Data + experiment design (setup db)* **lcg-cp/grid-url-copy db from local to SE**

b. *Application* **edg-job-submit BioDCV.jdl (jdl file)**

c. *Resulting db:* **lcg-cp/grid-url-copy db (from SE to local)**

ITC irst
CENTRO PER LA RICERCA SCIENTIFICA E TECNOLOGICA



BioDCV: Running on Microarray data

Breast cancer microarray dataset

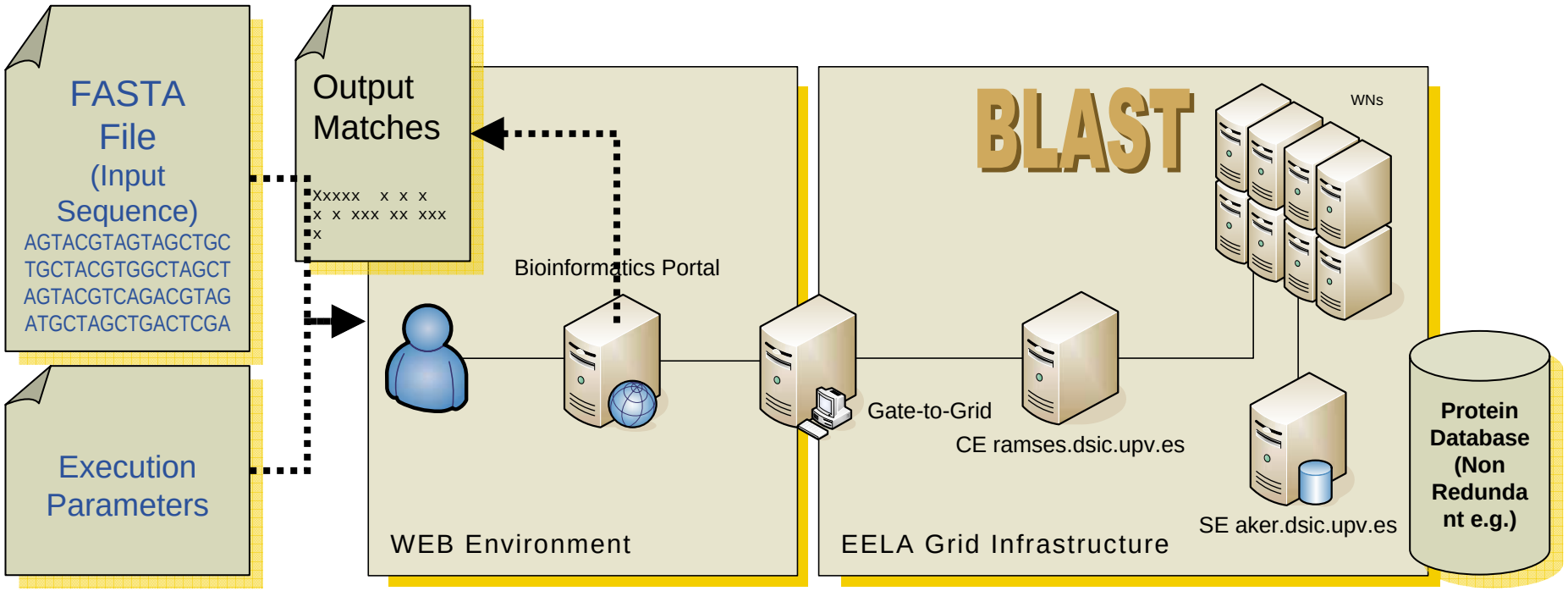
- **22215** genes and **183** samples
  4Mega footprint units (footprint = #features x #samples)
  Original work in (Sotiriou et al, J. Nat Canc Inst 2006)
  - September/October 2006
  - Used: 60 CPUs and about 40 Biomed sites.
  - 20 CPUs x 3 series (alternative machine learning models): RFE-Linear SVM, TR (Terminated Ramp) SVM, Correlation-aware RFE-SVM.
  - Failure: 5 % (3 jobs)
  - Running times (average over 20 runs):
    - Linear SVM ~ 5 hours
    - TR SVM ~ 8 hours
    - Correlation-Aware ~ 15 hours

PhyloJAVA

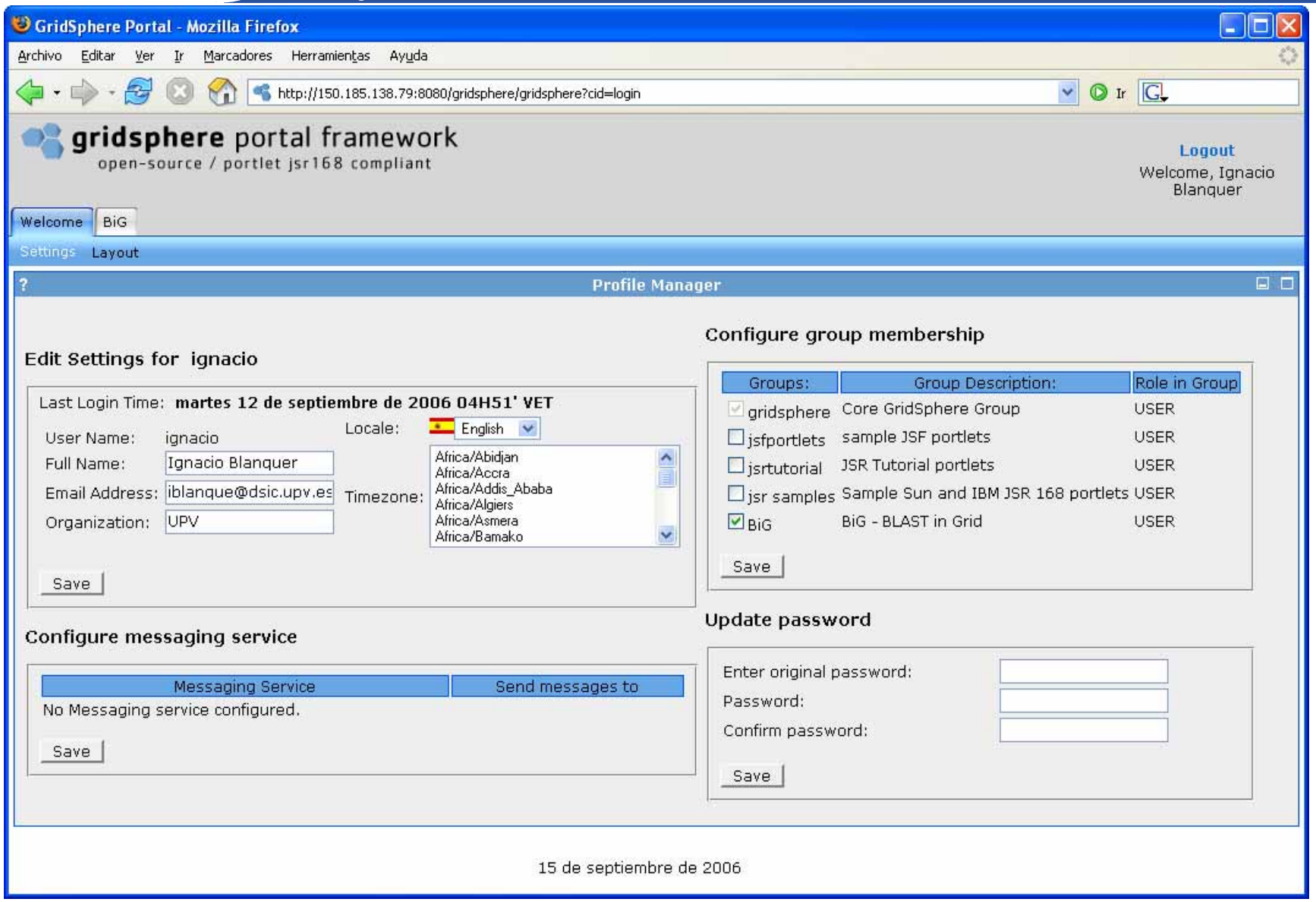
- **Phylojava is a client/server tool dedicated to phylogenetic tree reconstruction.**
  - This program allows phylogenetic tree inferences according to most usual methods (distance methods, maximum parsimony, maximum likelihood).
    - Phylogenetic trees are computed on a remote server, and are sent via internet to a graphical interface (the client) that allows the user to handle alignments and phylogenetic trees. The user therefore only has to install the graphical interface on his computer, and can submit tree reconstruction jobs on a remote server (EGEE grid or his computers).
- **Status:**
  - Porting to the EGEE grid.
- **Data sets need to be analysed**
  - about 300 sequences of more than 6000-characters-long each.
  - weeks of computation with the current bootstraping algorithms

BiG (1)

- **BLAST in Grids (BiG)**
  - Grid Interface to MPI Blast.
  - Access Through a Web Portal (http://portal-bio.ula.ve/).
  - Access to EELA Grid Through Gate-to-Grid Using a WSRF Interface.



BiG (2)


GridSphere Portal - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
http://150.185.138.79:8080/gridsphere/gridsphere?cid=login
gridsphere portal framework
open-source / portlet jsr168 compliant
Logout
Welcome, Ignacio Blanquer
Welcome
BiG
Settings
Layout
Profile Manager
Edit Settings for ignacio
Last Login Time: martes 12 de septiembre de 2006 04H51' VET
User Name: ignacio
Full Name: Ignacio Blanquer
Email Address: iblanque@dsic.upv.es
Organization: UPV
Locale: English
Timezone:
Africa/Abidjan
Africa/Accra
Africa/Addis_Ababa
Africa/Algiers
Africa/Asmera
Africa/Bamako
Save
Configure group membership
Groups: Group Description: Role in Group
gridsphere Core GridSphere Group USER
jsfportlets sample JSF portlets USER
jsrtutorial JSR Tutorial portlets USER
jsr samples Sample Sun and IBM JSR 168 portlets USER
BiG BiG - BLAST in Grid USER
Save
Configure messaging service
Messaging Service
Send messages to
No Messaging service configured.
Save
Update password
Enter original password:
Password:
Confirm password:
Save
15 de septiembre de 2006

SuperlinkOnline

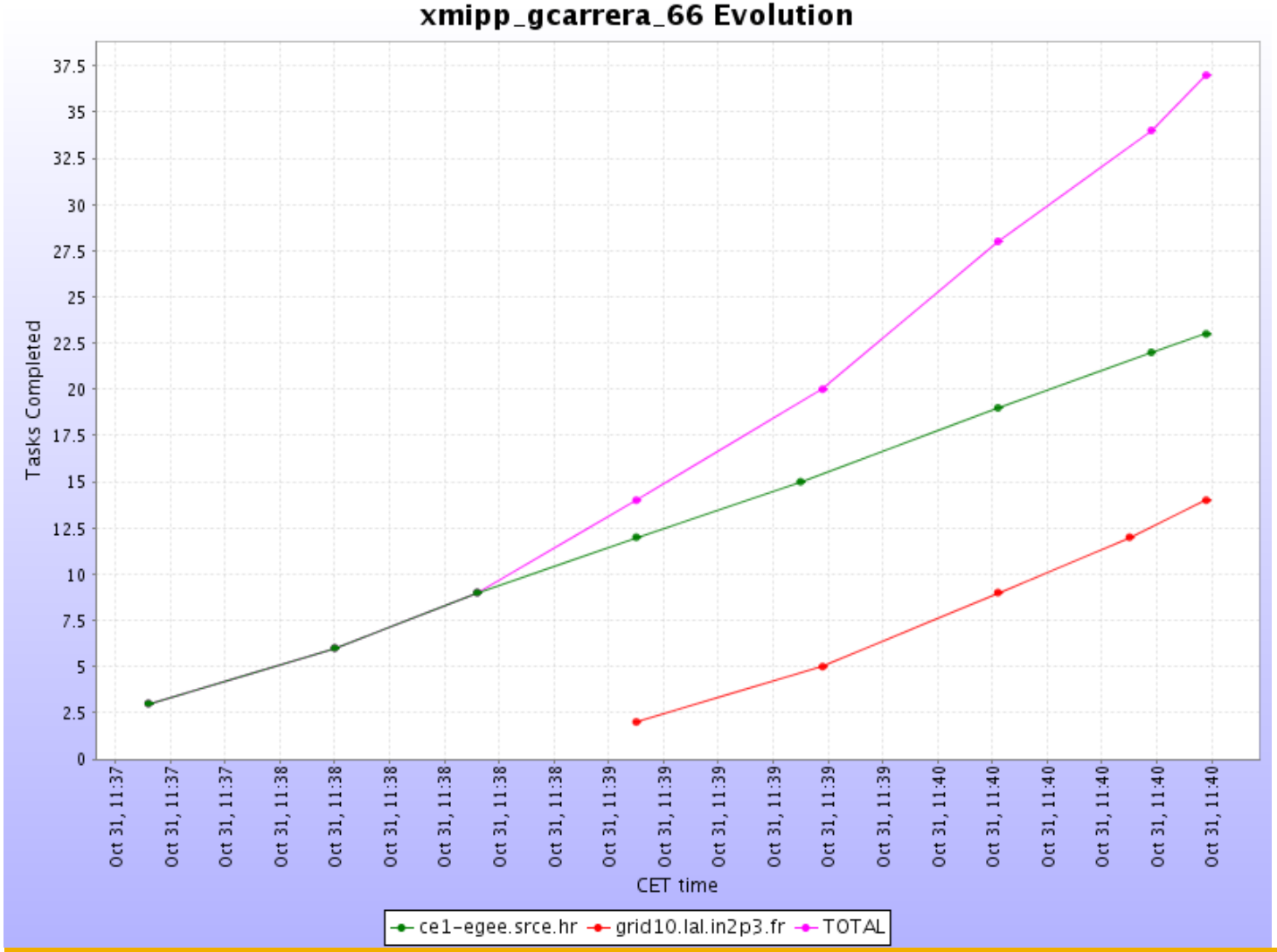
- **Superlink online : a tool for genetic linkage analysis**
  - Genetic Linkage Analysis is about hunting for disease-provoking genes.
- **Tasks are automatically divided into small pieces and executed simultaneously using many computers:**
    - Executable is pre-installed, very small I/O
    - From Level1 to L5 are from short tasks to very hard ones
    - EGEE Biomed VO addresses L4 “very hard tasks”
- **Statistics**
  - 7000 CPU-hours a day on ~3000 CPUS (Condor in Madison and Technion)
  - 20-40 runs daily
    - 1-3 runs of 10k jobs (15-30 min each)
    - 5-10 runs of 100-1000 jobs (15 min)
    - The rest are <30 jobs of up to 15 min
  - Workload will increase with new functionalities.

3DEM (1)

- **Selected by their user impact**
- **MLalign3D**
  - The combination of images in a 3D reconstruction requires:
    - That they represent projections of identical 3D objects.
    - That their relative orientations be known.
  - MLalign3D combines the tasks of
    - Classifying the images in homogeneous groups.
    - Aligning the images to obtain the best orientation.
  - MLalign3D employs a Maximum Likelihood method.
- **MLalign2D**
  - Is a similar to the 3D alignment case, only simpler

3DEM (2)

- **DIANE/MLalign2D jobs evolution**



MonAlisa plot
MLalign2D
37 Tasks
(each a subset of 10 images)

Grid for Bioinformatics

- **Very different applications …**
  - Different requirements and priorities
  - Different resources involved: hardware, software, human
  - Different Life science communities addressed
- **… but common requirements**
  - End-users don't care of the infrastructure !
  - Data
    - Deploying updatable databases
    - Security of biological data
  - Tools
    - Integrating numerous, complex programs: automatic procedure
    - Legacy application: grid-enabled without modification, SDJ, bundle, parallel job requiring MPI
    - Portal and user interfaces

- **Current major issues**
  - Workload Management
    - short job (< 5min): 2-3 min of overhead
    - bundle jobs: very long time submission (12h for 4,000 jobs)
  - Data management:
    - no tool in gLite to integrate database
  - Security:
    - data confidentiality, encryption
    - Portal certificate
    - management of long authentication (proxy)

Next Bioinformatics Meetings

- **EGEE Bioinformatics #3**
  - Valencia (Spain), **February 2007**, University of Valencia
  - Hosts: Vicente Hernandez, Ignacio Blanquer

- **EGEE Bioinformatics #4 (joint with EU Bioinfogrid)**
  - Varenna (Italy), **May 2007**
  - Host: Luciano Milanesi
- **EGEE Bioinformatics #5 (joint with SwissBioGrid)**
  - Lugano (Swiss), **Sept 2007**
  - Host: Peter Kunszt

- **Bioinformatics meetings are standing during 2 days**
  - One day for Internal EGEE bioinformatics activity report and discussion
  - One day for networking activity
    - with external applications and projects
    - workshop/tutorial about useful services: EGEE or 3rd-party ones

Next project meetings

- **EGEE User Forum 2**
  - Manchester, UK, May 9-11, 2007
  - Conjointly with OGF 20 (May 7-9)
  - Provide opportunities for an active dialogue between the EGEE project and its users (talks, demos, posters)
  - **http://www.eu-egee.org/uf2 , Call for Abstracts is open**

- **EGEE07**
  - Budapest, Hungary, 1-5 October 2007
  - Key European event dedicated to Grid technology: EGEE annual conferences are regularly attended by a large international Grid community coming together to discuss a wide range of issues, the latest developments, and international co-operation, with the aim of driving forward world-class Grid technologies.

Conclusion

- **Bioinformatics community on EGEE Grid**
  - 10 Applications
    - In production: **Splatche, bioDCV**
    - Prototype: **GPS@**
    - Porting: **Large Scale Pathway, BiG, 3DEM**, …
  - Provide expertise to port applications
- **Benefit from EGEE grid, largest platform in production mode**
- **Collaboration with related projects: EU EMBRACE, EU EELA, EU BIOINFOGRID, SwissBioGrid.**
- **Open to new applications: contact us.**

Build the Bioinformatics Grid

- **Bioinformatics services for developers**
  - Internal: integrate data and tools as grid services,
  - External: powerful interfaces: *e.g.* Web Services
- **High-level interfaces for end-users**
  - User-friendly: Web Portal for biologists, physicians
  - Efficient: integrated data and tools
- **Powerful interface to display Grid-scale results**
  - Thousands to millions of bioinformatics jobs
  - Graphical and Data-mining tools